AI at DOE & Argonne

Ian Foster
Argonne National Laboratory & The University of Chicago
foster@anl.gov

March 22 2023

AI4SES: AI 4 Science, Energy, & Security

Goal: Construct a report and plan that outlines and **makes the case for a 10-20 year program-project that enables the creation of a world leading capability in AI for DOE mission spaces**

- Identify directions, approaches, and where possible specific challenge problems, that should be pursued
- Identify the program scale needed to make progress
- Provide the "core content" that will help in forming budget requests and overall program approach
- Make the case for what needs to be done and why

FOCUS: Leadership AI for DOE mission needs

Scientific discovery, user facilities, energy research, environment and national security

Leverages relevant DOE assets

- Exascale class computing
- Exascale class data infrastructure
- Large-scale Experimental Facilities
- Large-scale Scientific Simulation Capabilities
- Interdisciplinary teams

Aiming for transformation of DOE research

- 1,300+ researchers participated in four town halls during summer 2019 and summer 2022: Modeled after exascale town halls in 2007-2009
- A DOE major initiative recommended in August 2020 by subcommittee of department’s Advanced Scientific Computing Advisory Committee
- Broad opportunities in AI
  - Biology, climate, chemistry, materials, physics, cosmology, nanoscience, fusion
  - Energy and national security
  - Integration with scientific facilities

Priority roles for AI in science, energy, & security

AI for advanced properties inference and inverse design

Energy storage, proteins, polymers

AI and robotics for autonomous discovery

Biology, chemistry, materials, photon and neutron sources

AI-based surrogates for high-performance computing

Climate ensembles, quantum chemistry, cosmology, effective zettascale on exascale

AI for software engineering and programming

Code translation, optimization, quantum compilation, algorithms

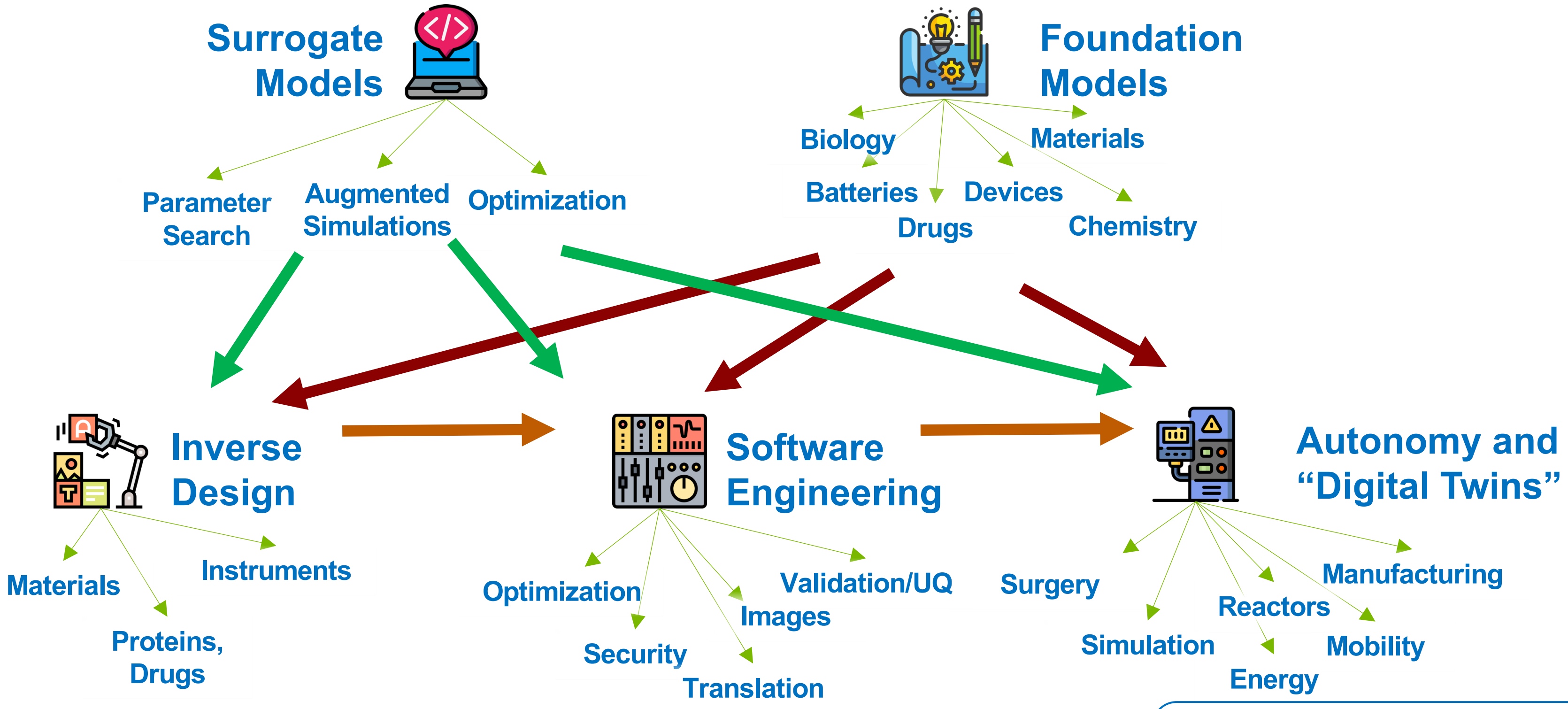
AI for prediction and control of complex engineered systems

Accelerators, buildings, cities, reactors, power grids, networks

Foundation AI for scientific knowledge

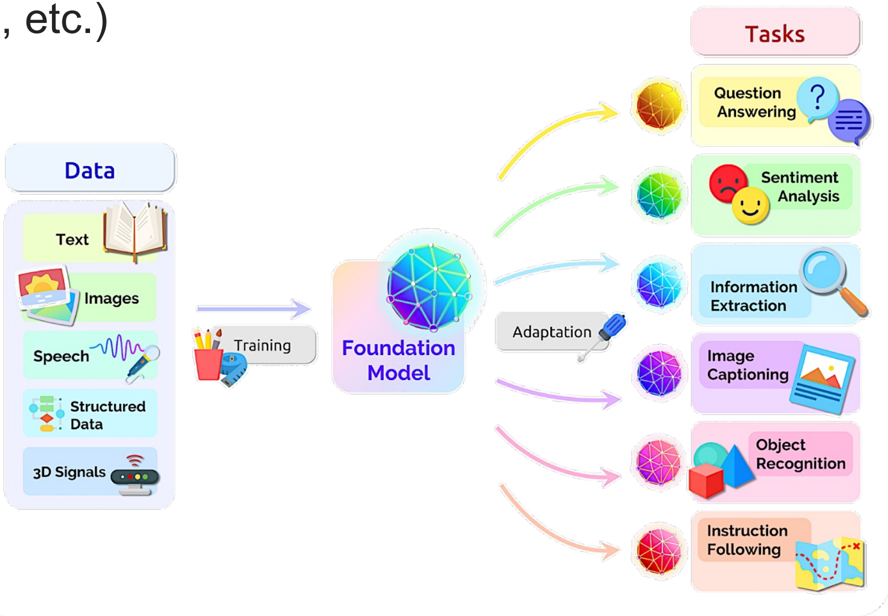
Hypothesis formation, math theory and modeling synthesis

Major emerging AI capabilities



Example: Foundation models for science

- Foundation models (LLMs, VLMs, etc.) are single large-scale models that have been pretrained in self supervised mode on large datasets from many sources (text, papers, datasets, code, molecules, etc.)
- Models are used in a “generative” fashion to compute “completions” in response to ”prompts”
- They are often wrapped in additional tools to clean up and filter outputs to improve the human interaction experience (e.g., ChatGPT)
- They are remarkably flexible and exhibit emergent behaviors at scale (e.g., spontaneously complete tasks they were not trained explicitly to do, such as translate between languages, or summarize text)
- Several efforts underway in DOE labs to build Foundation Models for science (e.g., 9 Yards at ORNL, AuroraGPT at Argonne)

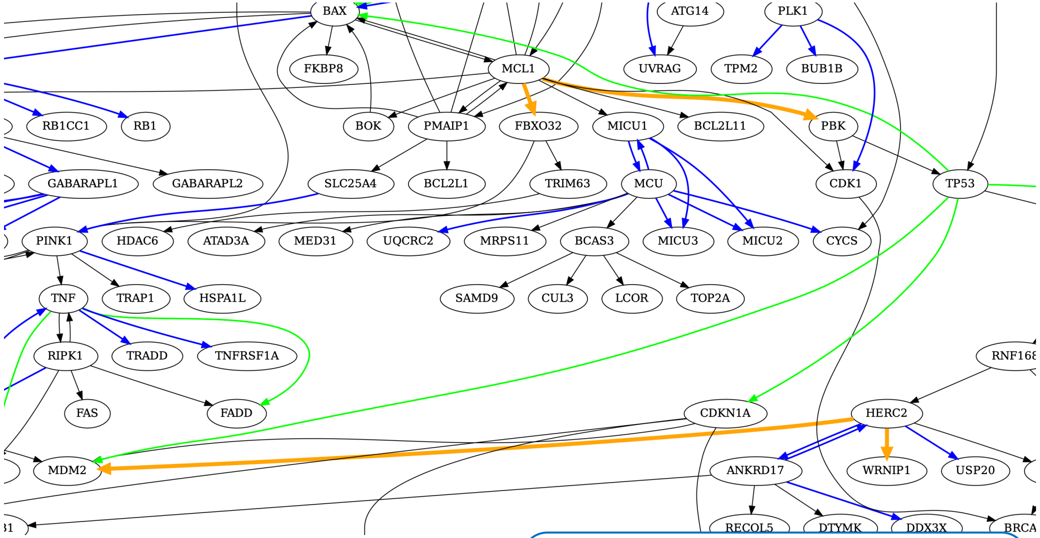


Foundation models for science — Opportunities

- FMs can **summarize and distill knowledge** – extract information from million of papers into compact computing representation – protein-protein interaction networks, materials compositions, code kernels, protein sequences, etc.
- FMs can **synthesize – combine information from multiple sources** – generate small programs for specific tasks – quantum computing programs using QISkit & Cirq, derivations for applied physics, code for visualization and animation, etc.
- FMs can **generate plans, solve logic problems and write experimental protocols** for robots – powering self-driving labs, generate strategies for problem solving, and planning for testing hypotheses
- FMs, with additional research, **may be able to generate hypotheses** to be tested and new theories for exploration – a full-time scientific assistant that learns from across all of DOE science

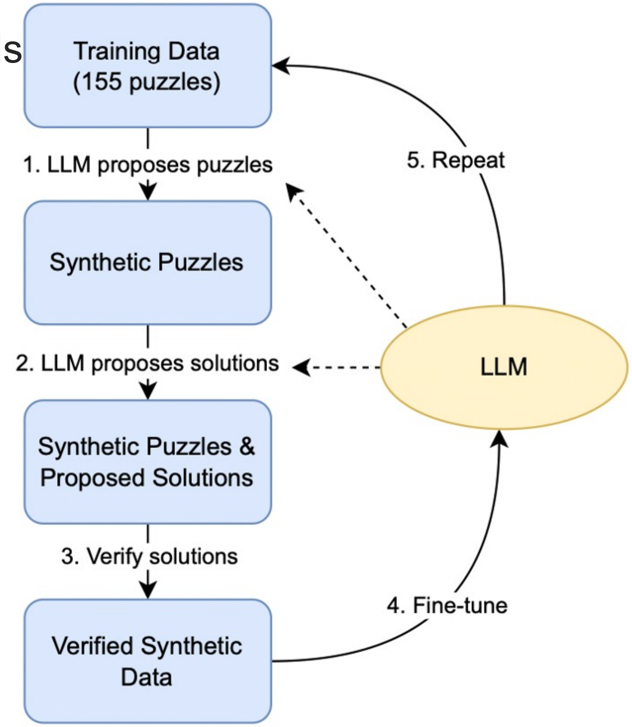
Foundation models — Impacts now and future

- Dramatically increased coding productivity (2x-3x has been demonstrated)
- Via APIs and remote access, extract in one weekend what would have taken months or years to do via traditional curation (PPI network reconstruction)
- Generate protein sequences for given purpose (function, interaction)
- Generate materials compositions that yield desired properties
- Given raw experiment data, generate paper summary, tables, figures
- Given conjectures and corollaries, generate a fully detailed proof
- Translate codes between languages
- Optimize code loops for GPUs
- Many, many others …



AI for software and programming

- LLMs specifically developed as coding assistants and coding aids have been developed (codex, palm-coder, etc.)
- Models are trained on large bodies of code (GitHub, etc.) using self-supervised MLM training schemes
- Models can be improved by boosting, generating random code against a simple set of random specifications and incorporating that code that correctly implements the spec
- These models can generate code, translate code, debug code and document code
- Recent systems can also uncompile code and translate binaries
- Current estimates are that for developers using these tools that ~40% of the code that is produced can be written by the LLMs
- Code generated is sometimes not correct, but if used as an assistant its usually quickly fixed
- Models are naturally modular with contexts (windows for training and generation) in the 4K to 32K tokens

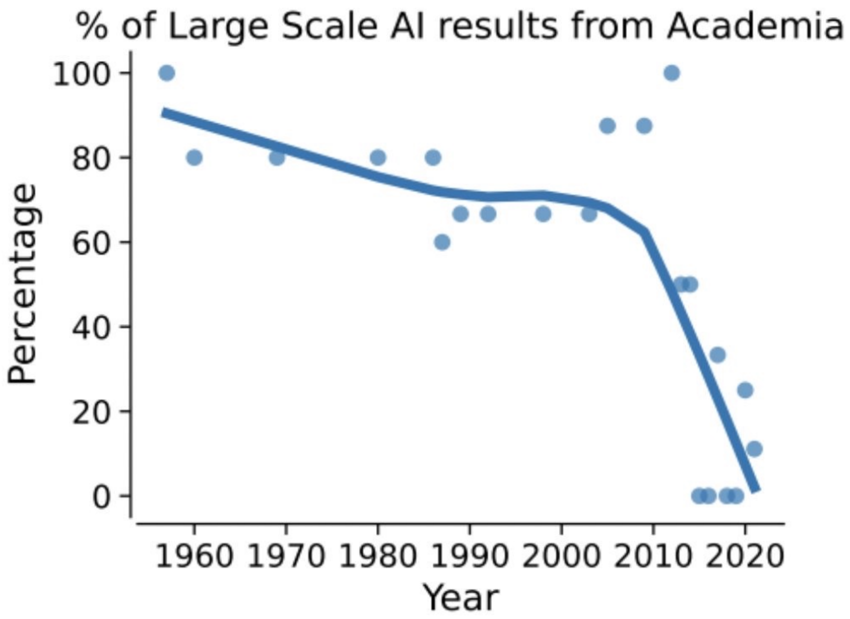


AI for software and programming — Opportunities

- Much of DOE science and technology research involves coding
- It has been estimated that DOE has more than 1 Billion LOC across the complex, most of which is not under active maintenance
- ECP investments resulted in 78 applications codes and over other 100 software projects being modernized and migrated to Exascale platforms and GPUs (estimated at 10-20 Million LOC)
- Build a FM for DOE scientific coding that knows about DOE code base
- Migrate codes to GPUs and future architectures
- Update codes to modern language versions (e.g. Python 3)
- Automatically document codes
- Improve performance through high-level code rewriting and parallelism
- Library interface porting to new versions
- Generate scientific codes from natural language descriptions

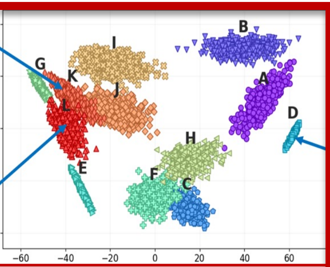
Thinking at scale is very important

- DOE and the laboratories were created to work on large-scale things in an interdisciplinary way
- Scale is part of what differentiates the labs from universities
- Leading edge research in AI today is dominated by large-scale groups and teams from industry.
- Large teams of people ~1000 per major AI research group
- Collections of projects organized around AI approaches with long horizons
- Serious software development effort in tools and software
- Access to vast computing resources
- Access to vast datasets

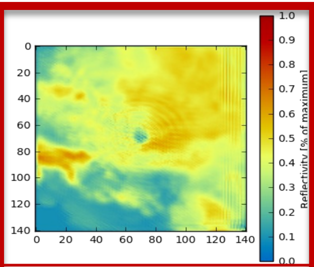


https://arxiv.org/pdf/2202.07785.pdf

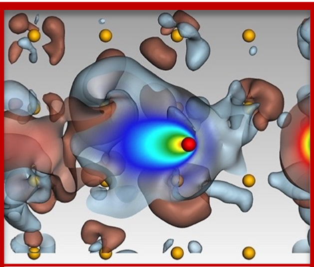
Sample of (out of 100+) ML/AI projects underway at Argonne



Reduced order modeling of laser sintering

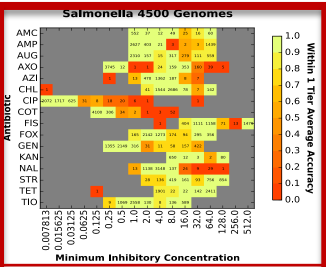


Nowcasting with convolutional LSTMs

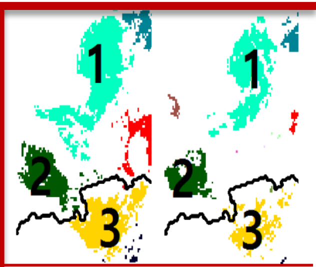


Prediction of radiation stopping power

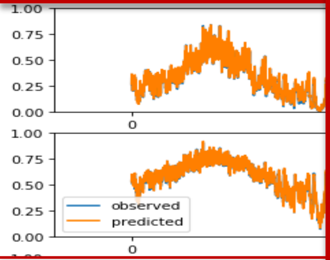
Strong and weak lensing in sky survey data



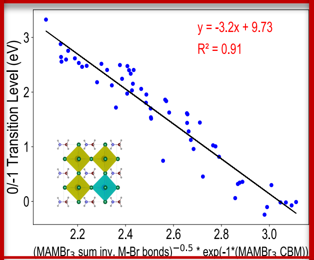
Prediction of antimicrobial resistance phenotypes



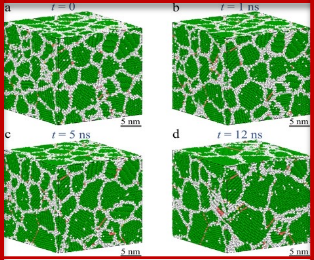
Identification and tracking of storms



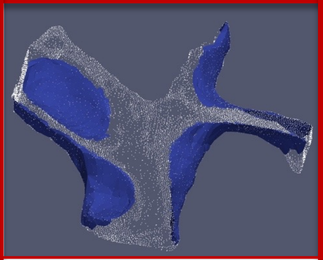
Efficient climate model emulators



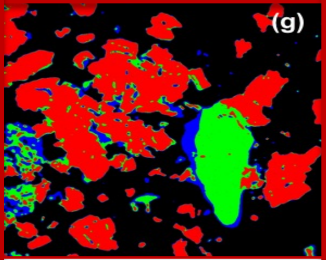
Defect-level prediction in seminconductors



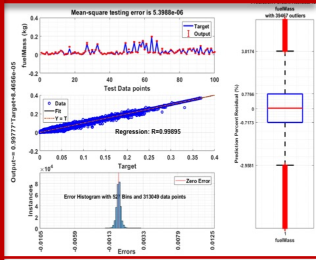
Structure-property-process triangle in additive manufact...



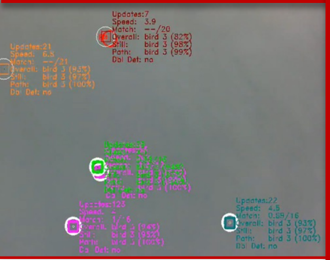
Parameter extraction in atom probe tomography



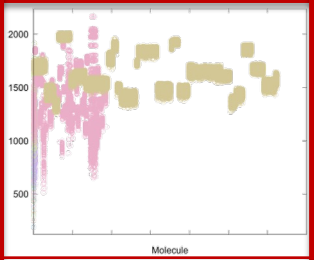
Learning for dynamic sampling in spectroscopy



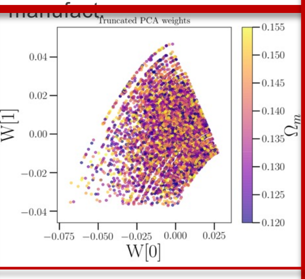
Vehicle energy consumption prediction



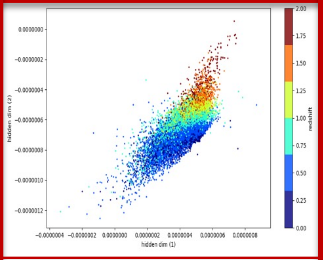
Flying object detector for edge deployment



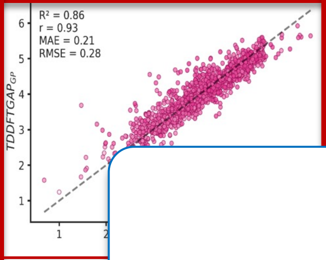
Discovery of new energy storage materials



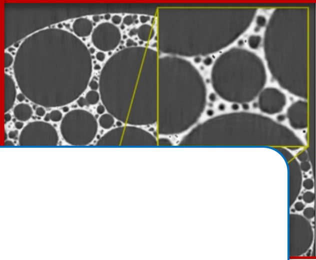
Cosmic Microwave Background emulation



Photometric red shift estimation



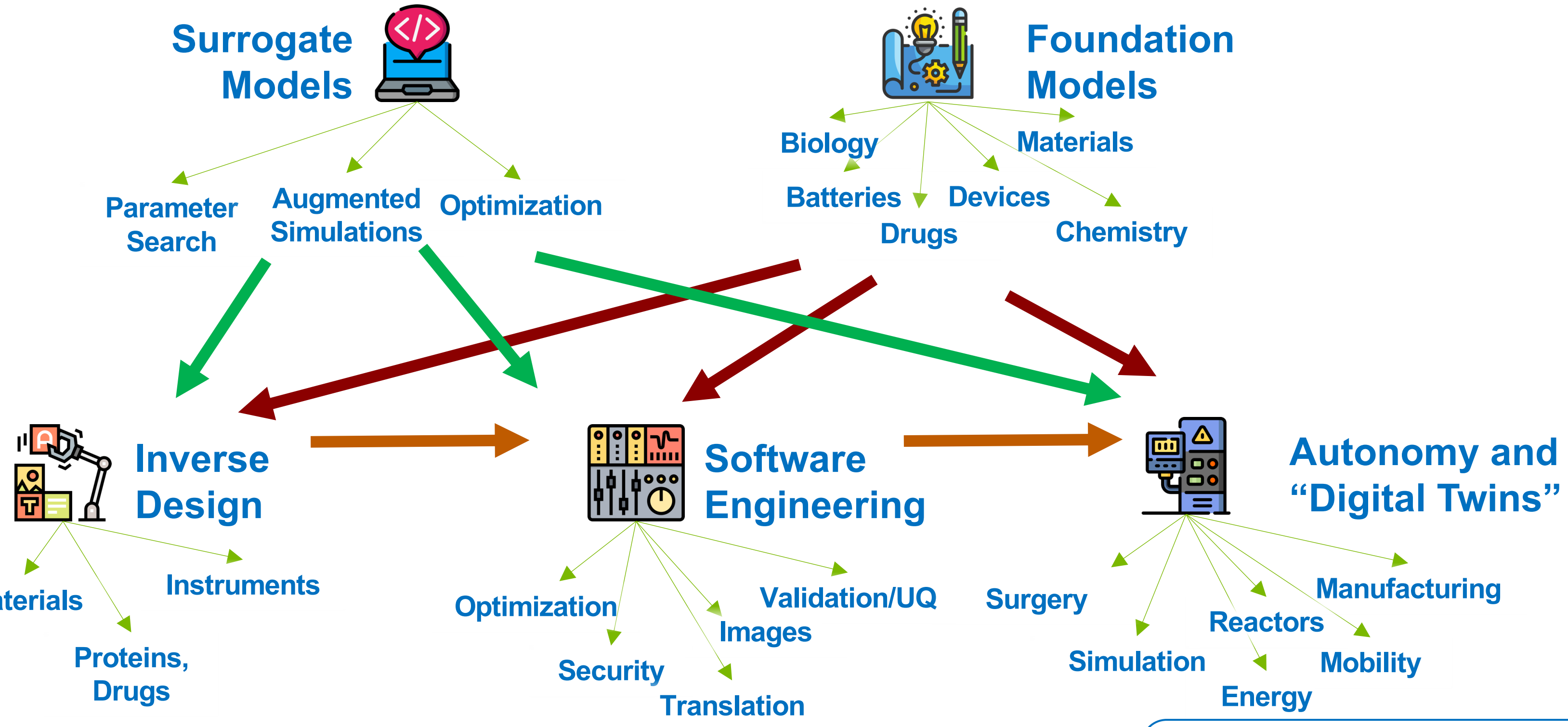
New ma... solar cells



tomographic images

Argonne’s Aurora System > 60,000 Intel GPUs: Science Starts in 2023

Preparing for AI at scale



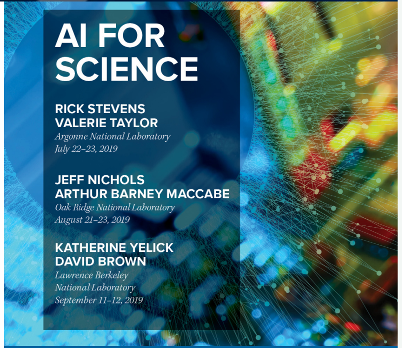
AI at Fermilab

Nhan Tran
March 22, 2023

Closed caption box size

Vision for HEP and AI

Today: how can HEP benefit **from national initiatives** and how can our nation benefit **from AI/ML in HEP**?



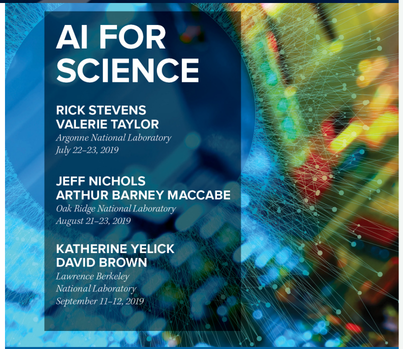
https://science.osti.gov/Initiatives/AI/

see also https://www.nsf.gov/cise/ai.jsp

Vision for HEP and AI

Ben Nachman at P5 LBNL Town Hall

Today: how can HEP benefit **from national initiatives** and how can our nation benefit **from AI/ML in HEP**?



https://science.osti.gov/Initiatives/AI/

see also https://www.nsf.gov/cise/ai.jsp

AI for physics, physics for AI

Builds **diverse, inclusive communities; assemble multi-disciplinary collaborations** around cross-cutting challenges

FNAL AI project office coordinating activities spanning the scientific directorates

Motivation

HEP builds and operates the most complex devices in science
AI is a **pervasive force multiplier** that can **enable transformative scientific capabilities**

Physics-inspired data & models

Robust & generalizable learning

"Fast" & efficient algorithms

Deeper insights & better performance

Accelerate time-to-physics

Improved efficiency and autonomous operations

Motivation

HEP builds and operates the most complex devices in science

AI is a **pervasive force multiplier** that can **enable transformative scientific capabilities**

Fermilab unique strength on real-time AI for accelerating HEP science

Motivation

HEP builds and operates the most complex devices in science

AI is a **pervasive force multiplier** that can **enable transformative scientific capabilities**

Fermilab unique strength on real-time AI for accelerating HEP science

Algorithms for HEP science

Physics-inspired data & models; Robust & generalizable learning; Fast and efficient algorithms

Intelligent sensing and real-time processing

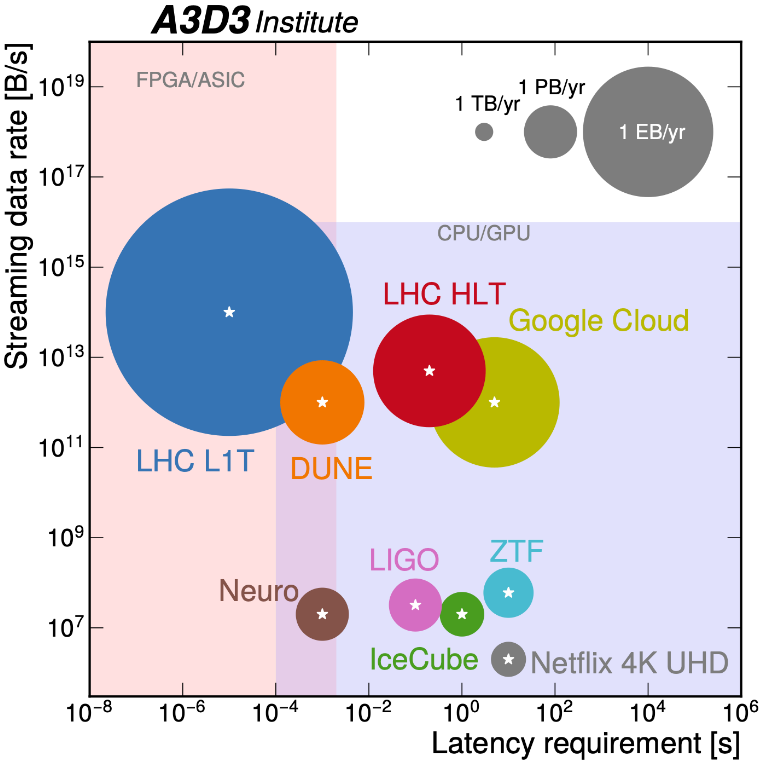
High performance and throughput compute

Operations, controls, analysis

Real-time & Fast ML

https://a3d3.ai/about.html
fastmachinelearning.org

Applications and Techniques for Fast Machine Learning in Science
https://doi.org/10.3389/fdata.2022.787421



Fusing powerful ML techniques with experimental design decreases the "time to science" and can range from embedding real-time feature extraction to be as close as possible to the **sensor all the way to large-scale ML acceleration** across distributed grid computing datacenters. The overarching theme is to **lower the barrier to advanced ML techniques and implementations** to make large strides in experimental capabilities across many seemingly different scientific applications. Efficient solutions **require collaboration** between domain experts, machine learning researchers, and computer architecture designers...

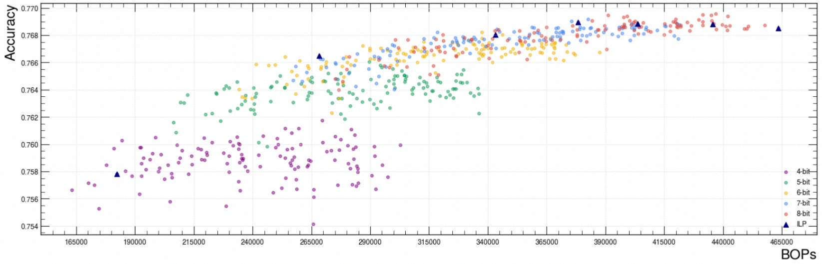
Quantization-aware pruning, arXiv:2102.11289
QONNX, arXiv:2206.07527
An end-to-end codesign workflow of Hessian-aware quantized neural networks for FPGAs and ASICs
Quantized Distilled Autoencoder Model for 4D Transmission Edge Microscopy

Fast and efficient algorithms

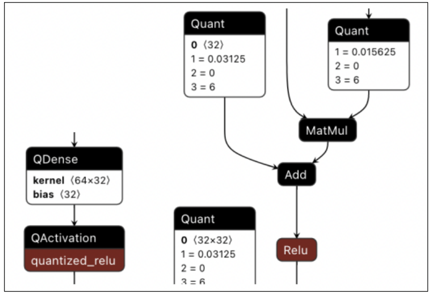
AI is data and energy *hungry*

- **Efficient and Robust AI**: very important for scientific sensing/compute
  - Broad applications, HEP and beyond
- Building techniques for wider scientific and industry communities
- **Core research** into:
  - quantization, sparsity,
  - multi-objective optimization
  - edge AI fault tolerance and robustness
  - DOE HEP project on efficient algorithms from inductive bias (physics-inspired)

Hessian-aware quantization solver more efficient than brute-force design

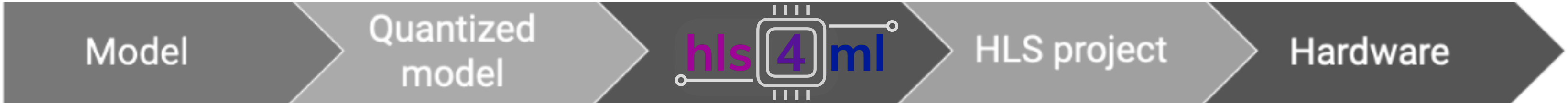


Industry/community standards for representing quantized neural networks



Embedded systems with HW-SW codesign

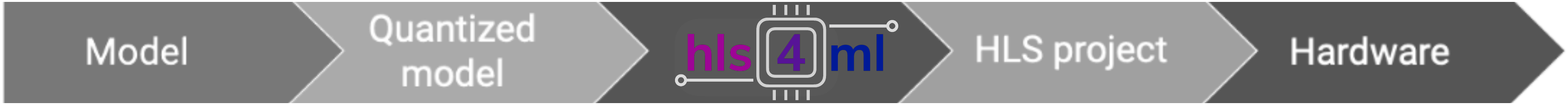
hls4ml, [JINST 13 P07027 (2018)](#)
https://fastmachinelearning.org/hls4ml
DUNE SNB, [TWEPP/IEEE NSS](#)
Reconfigurable ASCI, [IEEE TNS](#)



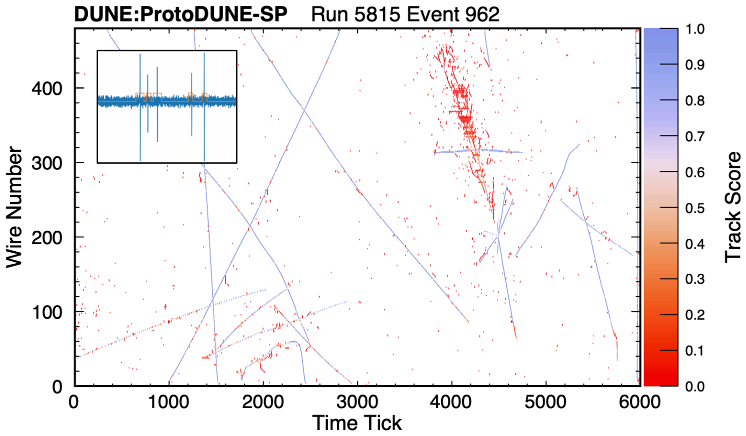
Strong synergy with microelectronics technologies

Embedded systems with HW-SW codesign

hls4ml, JINST 13 P07027 (2018)
https://fastmachinelearning.org/hls4ml
DUNE SNB, TWEPP/IEEE NSS
Reconfigurable ASCI, IEEE TNS



Extracting low energy neutrino signals per wire

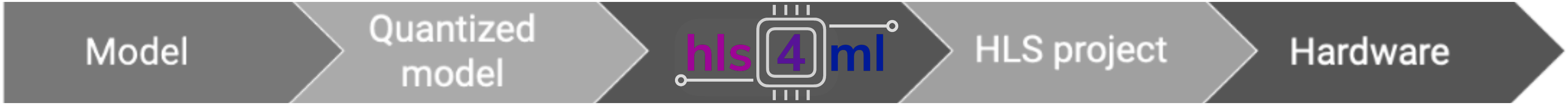


DUNE Supernova detection & multi-messenger astronomy

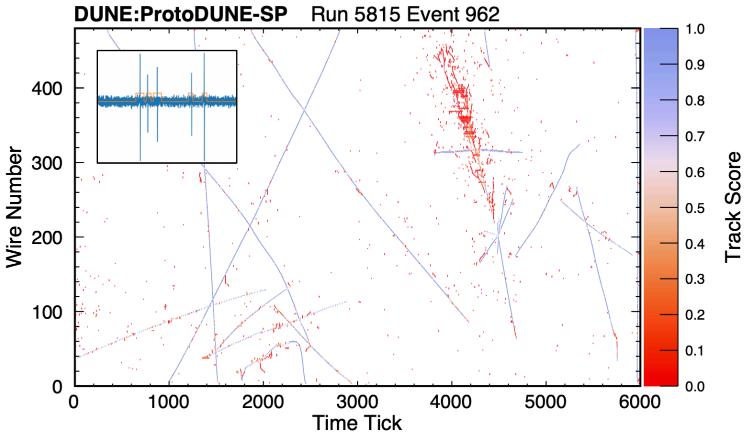
 Strong synergy with microelectronics technologies

Embedded systems with HW-SW codesign

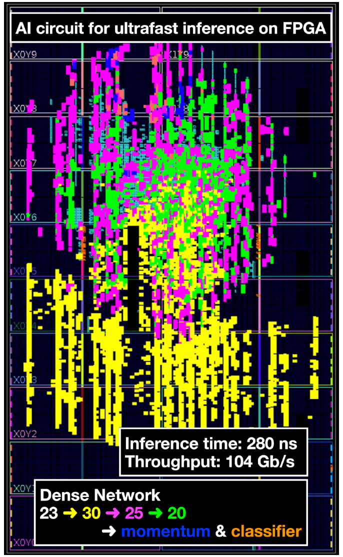
hls4ml, JINST 13 P07027 (2018)
https://fastmachinelearning.org/hls4ml
DUNE SNB, TWEPP/IEEE NSS
Reconfigurable ASCI, IEEE TNS



Extracting low energy neutrino signals per wire



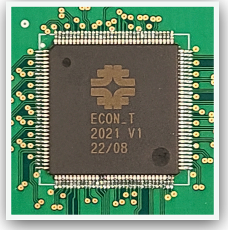
DUNE Supernova detection & multi-messenger astronomy



LHC Trigger - FPGA/ASIC

- New Run 3 algorithms *in hardware* for displaced muons and anomaly detection
- Several algorithms under investigation for HL-LHC trigger

- First modern AI algorithm in ASIC for CMS high granularity calorimeter
- Silicon-proven for functionality and radiation hardness
- R&D towards on-sensor pixel readout



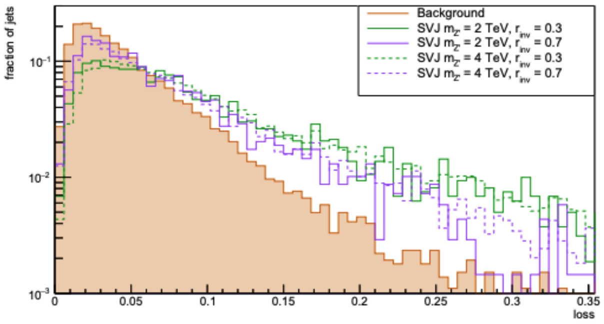
Strong synergy with microelectronics technologies

Robust and Physics-inspired AI

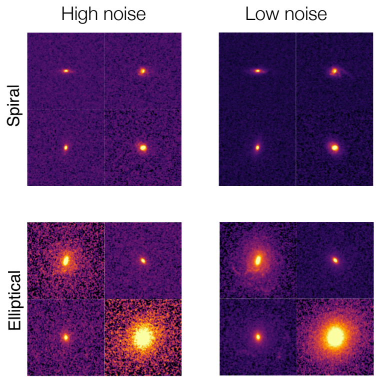
JHEP 02 (2022) 074
arXiv: 2107.02157
Nature Machine Intelligence 4, 154 (2022)
arXiv: 2110.08508
EPJC
DUDA, 2022 Neurips Workshop
arXiv:2102.06976
deepskieslab.com

Robust learning paramount for real-time sensing and controls

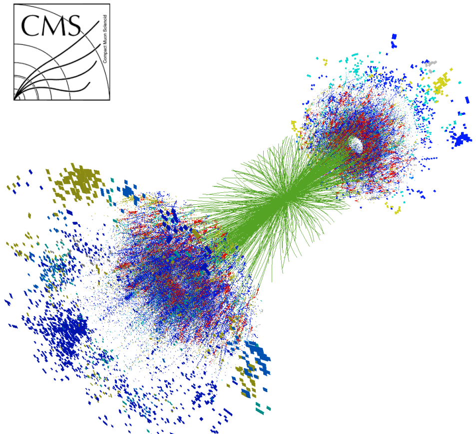
Physics-inspired models key for robustness and efficiency



Anomaly detection for monitoring, controls, and discovery



Domain adaptation to adjust to new datasets and conditions

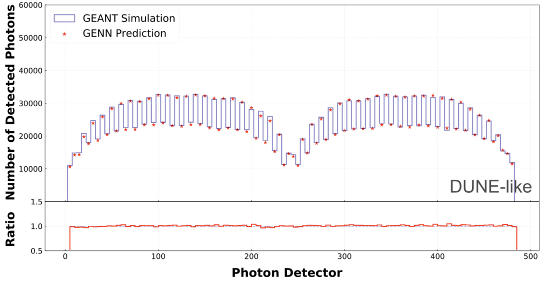


Optimal physics representations & architectures

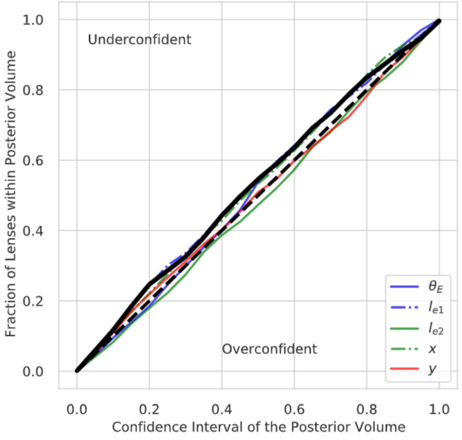
Robust and Physics-inspired AI

JHEP 02 (2022) 074
arXiv: 2107.02157
Nature Machine Intelligence 4, 154 (2022)
arXiv: 2110.08508
EPJC
DUDA, 2022 Neurips Workshop
arXiv:2102.06976
deepskieslab.com

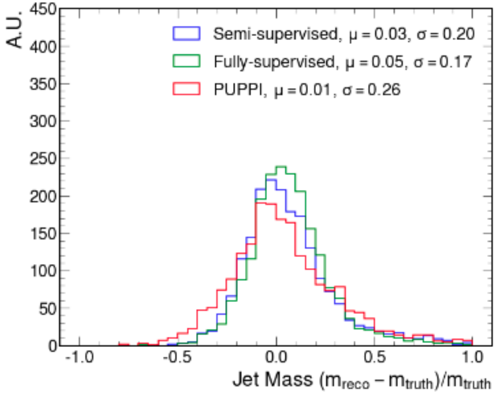
Robust learning paramount for real-time sensing and controls
Physics-inspired models key for robustness and efficiency



AI-accelerated simulation based on physics modeling



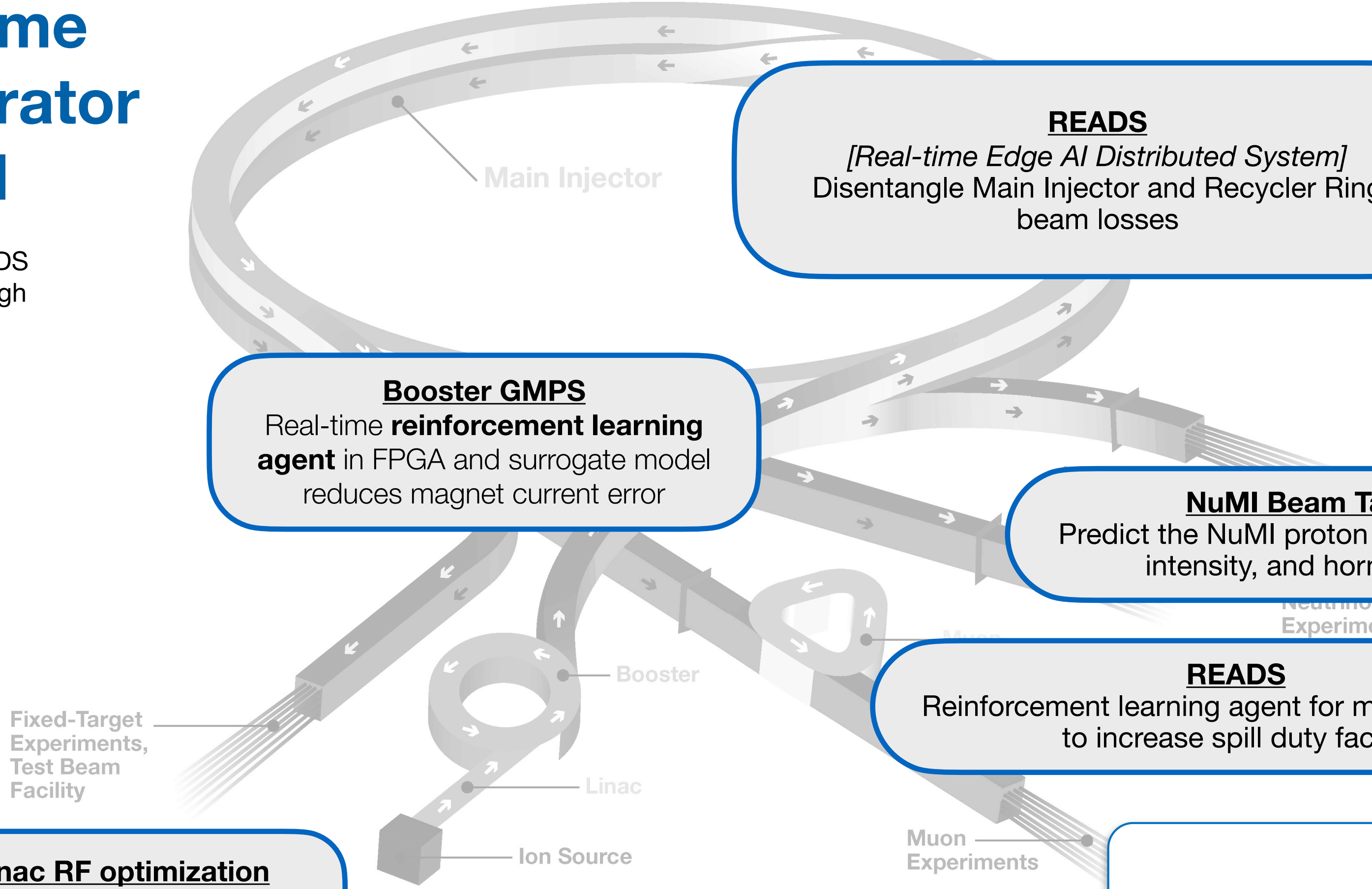
High-dimensional data reduction (likelihood-free) requires uncertainty quantification (UQ)



Semi-/self-supervised algorithms, reduce reliance on simulation

Real-time accelerator control

Support for READS and LINAC through DOE user facility grants



<u>READS</u>
[Real-time Edge AI Distributed System]
Disentangle Main Injector and Recycler Ring beam losses

<u>Booster GMPS</u>
Real-time **reinforcement learning agent** in FPGA and surrogate model reduces magnet current error

<u>NuMI Beam Target</u>
Predict the NuMI proton beam position, intensity, and horn current

<u>READS</u>
Reinforcement learning agent for mu2e slow spill to increase spill duty factor

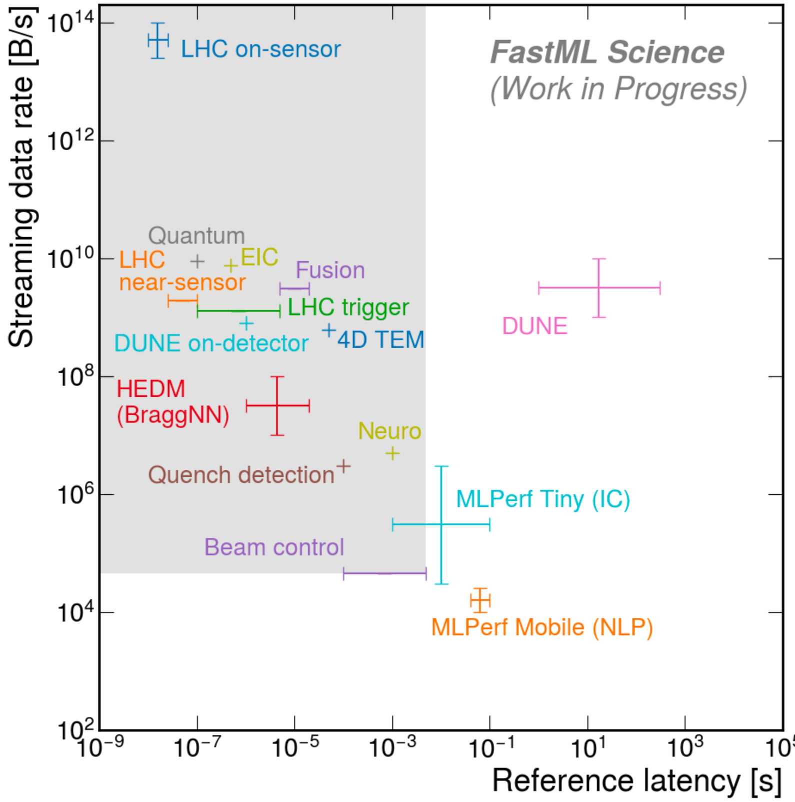
<u>Linac RF optimization</u>
to keep beam energy constant and minimize emittance

HEP for real-time AI

Nexus for developing Fast ML benchmarks across science

Grand Challenges spur innovation

- *LHC*: all sub-detectors analyzing data at 40 MHz
- *DUNE:* expansive (non-)accelerator ν program (solar, supernova, proton decay, ββ decay)
- *Accelerator controls* with adaptive online agents and digital twin
- Science: Quantum, Magnets, Fusion, Neuroscience, Nuclear, Material sciences, etc.
- Industry: Internet-of-Things, AVs, manufacturing



HEP for real-time AI

Nexus for developing Fast ML benchmarks across science

Grand Challenges spur innovation

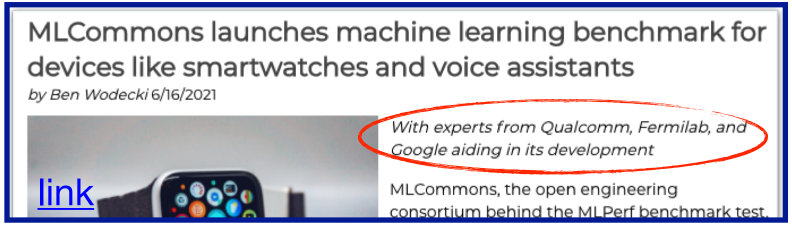
- *LHC*: all sub-detectors analyzing data at 40 MHz
- *DUNE:* expansive (non-)accelerator ν program (solar, supernova, proton decay, ββ decay)
- *Accelerator controls* with adaptive online agents and digital twin
- Science: Quantum, Magnets, Fusion, Neuroscience, Nuclear, Material sciences, etc.
- Industry: Internet-of-Things, AVs, manufacturing

Partnerships multidisciplinary collaboration with industry, academia, and other scientific domains

SLAC National Accelerator Laboratory
Sandia National Laboratories
Argonne National Laboratory
Berkeley Lab
Pacific Northwest National Laboratory
Oak Ridge National Laboratory
Los Alamos National Laboratory
Jefferson Lab

AMD
Mentor A Siemens Business
Tiny ML
ML Commons

+ many university partners and others!


MLCommons launches machine learning benchmark for devices like smartwatches and voice assistants

by Ben Wodecki 6/16/2021

With experts from Qualcomm, Fermilab, and Google aiding in its development

MLCommons, the open engineering consortium behind the MLPerf benchmark test.

link

Outlook

Artificial Intelligence is a pervasive force multiplier for physics

Transformative scientific capabilities from physics grand challenges

Increased investment in diverse collaborations for AI for particle physics has, and will continue, to bring new technologies to other scientific domains and industry